C2I-CASITI-VTG: CONFIDENCE-TO-IOU CASCADED INFERENCE-TIME INTERVENTION FOR VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding (VTG) aims to localize the temporal segment in an untrimmed video that corresponds to a natural-language query. Recent inference-time intervention (ITI) methods improve model behavior by steering internal activations. A natural VTG strategy is to subtract the mean activation of low-IoU samples from that of high-IoU samples, encouraging better-localized predictions. However, high IoU does not necessarily indicate that the underlying activation reliably represents successful localization. IoU and model confidence are sometimes misaligned: predictions can have high IoU with low confidence, or high confidence with poor overlap with the ground truth. Thus, IoU-only contrasts may mix inconsistent internal behaviors and produce noisy steering directions. We propose a **confidence-to-IoU cascaded ITI framework for VTG (C2I-CasITI-VTG)**. First, confidence-based ITI modifies the model's confidence behavior; IoU-based ITI then builds on these confidence-aligned representations to improve localization. Technically, the interventions cannot be naively combined as independently estimated steering vectors may point in incompatible directions and their superposition may cancel useful components. We therefore apply confidence-based ITI during prefill stage to reshape the subsequent decoding activations toward a confidence-conditioned representation space, from which we extract the IoU steering direction. We demonstrate that ITI can be cascaded to steer model behavior sequentially along complementary directions. Experiments on Charades-STA, ActivityNet-Captions, and QVHighlights show consistent gains across frozen MLLMs. On Charades-STA, our method improves Qwen3.5-9B from 53.93% to 57.47% mIoU and from 32.72% to 41.16% R1@0.7.

1 INTRODUCTION

Video temporal grounding (VTG) aims to identify the temporal interval in an untrimmed video that best matches a natural-language query (Gao et al., 2017; Hendricks et al., 2018). Recent multimodal large language models (MLLMs) have shown promising training-free video reasoning and temporal understanding capabilities by jointly modeling video content and language instructions (Maaz et al., 2024; Lin et al., 2023). Nevertheless, recent studies continue to report substantial challenges in predicting accurate temporal boundaries with video MLLMs, especially under strict overlap criteria(Wu et al., 2026; Xu et al., 2025; Yang et al., 2025). Fine-tuning can improve localization, but updating billions of parameters is computationally expensive and may compromise the general-purpose capabilities of the pretrained model. It is therefore desirable to improve VTG performance while keeping the MLLM frozen.

Inference-time intervention (ITI) (Li et al., 2023) provides a parameter-efficient means of doing so. ITI identifies a behavior-related direction in a model's activation space, typically by contrasting the mean hidden activations of positive and negative examples, and shifts activations along this direction at inference time. A natural application to VTG is to divide examples according to their intersection over union (IoU) with the ground-truth interval and construct an IoU direction by subtracting the mean activation of low-IoU examples from that of high-IoU examples. Injecting this direction should, in principle, steer the frozen model toward more accurate temporal boundaries.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

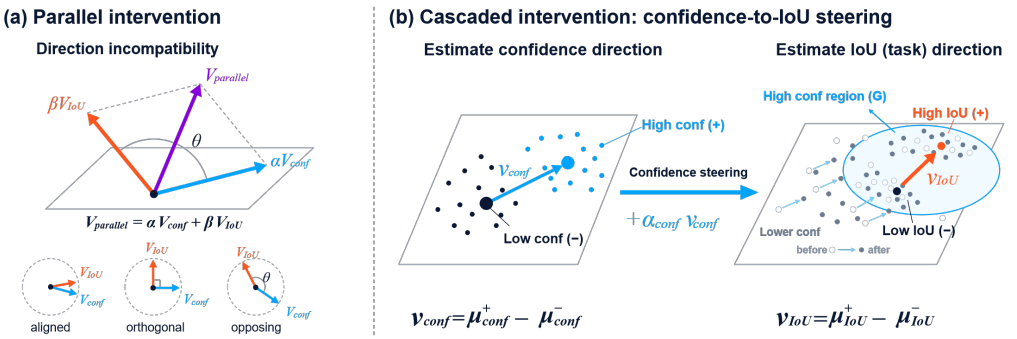


Figure 1: Motivation for confidence-to-IoU cascading. Independently learned confidence and IoU directions can be incompatible, while their weighted sum requires an additional balancing coefficient. Our method instead uses confidence to construct the representation space before constructing and deploying a single IoU direction.

However, high IoU does not necessarily mean that the activation underlying a prediction reliably represents successful localization. We observe a frequent mismatch between external localization quality and the model's internal certainty: a prediction can attain high IoU despite low confidence, whereas a highly confident prediction can overlap poorly with the ground truth. Consequently, an IoU-based partition may group together samples with substantially different internal behaviors. Directly averaging their activations entangles localization-related variation with confidence-related variation, producing a noisy IoU direction. This distinction is important because IoU and confidence describe different aspects of a prediction: IoU is an external, ground-truth-dependent measure of temporal overlap, while confidence reflects the model's internal belief in its own decoding decisions. Treating high IoU alone as evidence of a reliable internal state therefore limits IoU-based ITI.

An intuitive remedy is to estimate separate confidence and IoU steering vectors and combine them. Yet such a parallel intervention is not straightforward. Independently estimated directions may contain incompatible components (Fig. 1(a)), and their weighted superposition can weaken useful components or cause the two steering effects to interfere. Moreover, confidence and localization are not independent objectives: changing confidence-related activations early in inference alters the hidden-state on which the subsequent localization decision is based. A static sum of two vectors neither models this dependency nor exploits the ordered structure of autoregressive inference. It also introduces additional intervention coefficients whose balance can be sensitive across models and datasets.

To address these issues, we propose C2I-CasITI-VTG, a confidence-to-IoU cascaded inference-time intervention framework for video temporal grounding. Instead of treating confidence and IoU as parallel objectives, C2I-CasITI-VTG steers them sequentially. The first stage applies confidence-based ITI during prefill, reshaping the query-conditioned video hidden-state before interval decoding begins. This early intervention modifies the model's confidence behavior and allows its effect to propagate into the ensuing decoding trajectory. The second stage then applies IoU-based ITI during generation, operating on the confidence-aligned hidden-state to refine the predicted temporal boundaries. The confidence-to-IoU ordering is deliberate: confidence steering first regularizes the internal behavioral state from which localization is generated, enabling the subsequent IoU direction to target temporal quality with less confidence-related interference (Fig. 1(b)).

The central innovation of our method is thus not merely the use of two steering signals, but the discovery that ITI can be cascaded across inference stages to adjust model behavior sequentially along complementary directions. Existing ITI formulations predominantly steer a single behavioral axis or combine several independently learned directions in parallel. In contrast, our formulation makes one intervention establish the representation conditions under which the next intervention operates. This converts activation steering from a single-step vector shift into a structured multi-stage process and offers a general mechanism for coordinating dependent behavioral objectives without updating model parameters. For VTG, it provides a principled way to reconcile internal confidence with external localization quality.

We evaluate C2I-CasITI-VTG on Charades-STA (Gao et al., 2017), ActivityNet-Captions (Krishna et al., 2017), and QVHighlights (Lei et al., 2021) using multiple MLLMs. The method consistently improves temporal grounding across datasets and model families. On Charades-STA, it improves Qwen3.5-9B (Qwen Team, 2026) from 53.93% to 57.47% mIoU and from 32.72% to 41.16% R1@0.7, demonstrating particularly strong gains under a strict localization criterion. These results validate both the effectiveness of confidence-to-IoU intervention for VTG and the broader potential of cascaded ITI as a multi-directional activation-steering paradigm.

Our contributions are summarized as follows:

1. We reveal that direct IoU-based ITI for VTG is limited by a confidence–IoU mismatch: high localization quality does not necessarily correspond to a reliable internal activation state, causing IoU-only activation contrasts to entangle inconsistent behaviors.
2. We introduce C2I-CasITI-VTG, a confidence-to-IoU cascaded ITI framework that applies confidence steering during prefill and IoU steering during generation, exploiting their dependency and the temporal structure of autoregressive inference.
3. We establish a more general activation-steering principle: ITI can be cascaded across inference stages to steer a frozen model sequentially along complementary behavioral directions, rather than being restricted to a single direction or a parallel vector combination.
4. We demonstrate consistent improvements across three VTG benchmarks and multiple MLLMs, including gains of 3.54 mIoU points and 8.44 R1@0.7 points for Qwen3.5-9B on Charades-STA.

2 Related Work

Video temporal grounding. Video temporal grounding localizes a natural-language query in an untrimmed video through cross-modal matching, proposal modeling, span prediction, or joint moment–saliency prediction (Gao et al., 2017; Hendricks et al., 2018; Zhang et al., 2020b;a; Lei et al., 2021). Recent MLLMs improve temporal understanding through representation learning, post-training, or reinforcement learning (Deng et al., 2025; Li et al., 2025b; Wang et al., 2025a; Li et al., 2025a; Zhang et al., 2026). Our method is complementary: it improves localization at inference time with the backbone frozen.

Training-free temporal grounding. Frozen MLLMs can also be adapted through temporal evidence presentation and search. NumPro numbers video frames, Moment-GPT combines query reformulation with candidate generation and scoring, and VTimeCoT introduces visual temporal cues and chain-of-thought reasoning (Wu et al., 2025; Xu et al., 2025; Zhang et al., 2025). T2SGrid arranges neighboring frames into timestamped spatial grids, whereas HiTeA decomposes videos into temporal hierarchies (Guo et al., 2026; Xu et al., 2026). Here, NumPro and T2SGrid denote their training-free variants. These methods improve how temporal evidence is presented or searched. C2I-CasITI-VTG instead modifies the internal activations that process this evidence and generate temporal boundaries.

Activation steering and intervention composition. Inference-time intervention (ITI) controls model behavior through activation changes. ITI and Contrastive Activation Addition construct behavioral directions from activation contrasts (Li et al., 2023; Rimsky et al., 2024). Dynamic Multimodal Activation Steering selects context-dependent truthfulness vectors and combines them with visual-perception interventions, while AUSteer restricts steering to selected atomic units (Yin et al., 2026; Feng et al., 2026). These studies emphasize intervention context and granularity. For multiple objectives, MAT-Steer reduces attribute conflicts through selective intervention and sparsity and orthogonality constraints (Nguyen et al., 2025). PID Steering uses feedback to account for accumulated deviations and rapid activation changes (Nguyen et al., 2026). Our focus is the ordering and computational placement of complementary interventions within generative VTG.

Confidence-guided localization and cascaded ITI. Intersection over union (IoU) measures ground-truth overlap, whereas confidence reflects predictive behavior; higher confidence alone does not imply better calibration (Guo et al., 2017). URPA uses rollout variability to weight adaptation

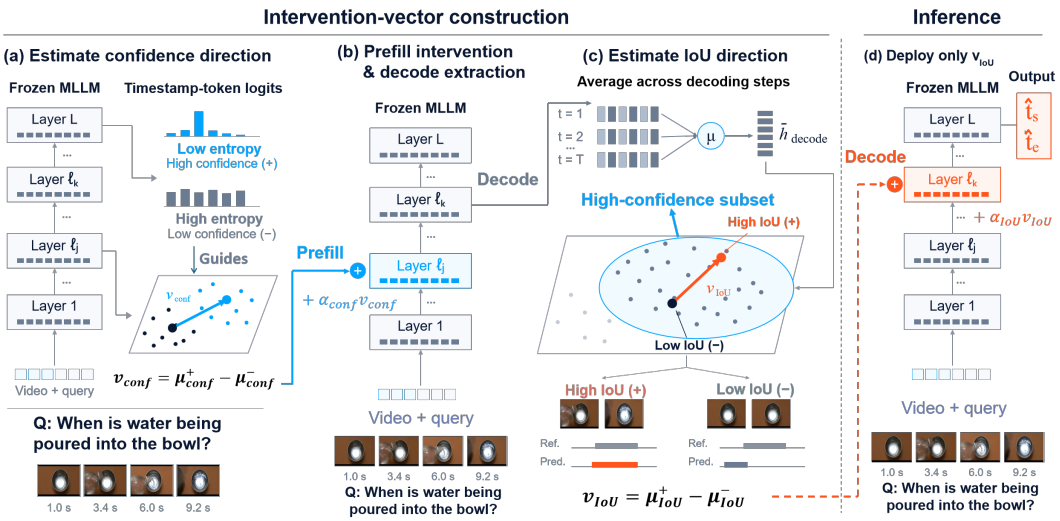


Figure 2: Overview of C2I-CasITI-VTG. (a)–(c) construct the intervention vectors from calibration examples. The model generates the start and end timestamps token by token; each generated token corresponds to one decoding step with its own logits. We use these logits to estimate confidence and obtain $\mathbf{v}_{\text{conf}}$ in (a), inject it during prefill in (b), and then use the resulting high-confidence representations to obtain $\mathbf{v}_{\text{IoU}}$ in (c). At inference time, only $\mathbf{v}_{\text{IoU}}$ is injected during generation (d).

rewards (Hu et al., 2025), while UG-Ground aggregates window-level affirmative–negative probability differences for training-free localization (Kim et al., 2026a). We investigate confidence as an activation-level intervention objective. Contrasting high- and low-IoU activations naturally targets localization quality, but may pool distinct confidence behaviors. Independently estimated confidence and IoU directions may also contain opposing components, making simultaneous addition insufficient to preserve their individual effects. C2I-CasITI-VTG cascades interventions during steering-vector construction: confidence-based ITI modifies prefill activations, and the IoU direction is then extracted from the resulting confidence-conditioned decoding activations. This ordering targets the dependency between hidden-state formation and interval generation.

3 Methodology

3.1 Problem Formulation and Preliminaries: Inference-Time Intervention

Given an untrimmed video V and a natural-language query Q, a frozen multimodal large language model (MLLM) autoregressively predicts a temporal interval $\hat{\tau} = (\hat{t}_s, \hat{t}_e)$, where $\hat{t}_s$ and $\hat{t}_e$ denote the predicted start and end timestamps. The ground-truth interval is $\tau = (t_s, t_e)$, and temporal localization quality is measured by $\text{IoU}(\hat{\tau}, \tau)$. During interval decoding, the model also produces a logit distribution over candidate timestamp tokens. These logits expose the model's expressed certainty and provide an internal behavioral signal that is distinct from the ground-truth-dependent IoU.

Inference-time intervention (ITI) modifies hidden activations of the model while leaving its parameters unchanged. Let $\mathbf{h}_+$ and $\mathbf{h}_-$ denote activations from positive and negative samples at a selected layer. A conventional ITI direction is the empirical mean difference

$$\mathbf{v} = \mathbb{E}[\mathbf{h}_+] - \mathbb{E}[\mathbf{h}_-], \tag{1}$$

where $\mathbf{h}_+, \mathbf{h}_- \in \mathbb{R}^d$ and d is the hidden dimension. At inference time, the activation is shifted according to

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha \mathbf{v}, \tag{2}$$

where ℓ is the intervention layer and α controls the intervention strength.

Our Motivation: Since confidence reflects the model's internal behavioral state, it should ideally be correlated with prediction accuracy; however, this correspondence is not always guaranteed.

Consequently, an IoU contrast in the original activation space captures localization differences but overlooks the importance of confidence as a signal of the model's internal behavior. C2I-CasITI-VTG addresses this dependency by constructing the localization direction only after confidence behavior has been internalized.

3.2 Stage I: Confidence Intervention during Prefill

Temporally Weighted Entropy for Confidence Quantification. Given a video and a natural-language query, the frozen MLLM generates a textual response specifying the start and end times of the video segment that it predicts matches the query. These two timestamps define the predicted temporal interval. For example, the response "from $3.6\,\mathrm{s}$ to $9.8\,\mathrm{s}$" specifies the interval $[3.6, 9.8]$ seconds, with 3.6 as the start timestamp and 9.8 as the end timestamp.

The response is generated autoregressively, one token at a time; each token-generation operation is a decoding step. A timestamp can therefore span multiple decoding steps. In the example above, suppose 3.6 is tokenized as 3, `.`, and 6. Generating these three tokens takes three decoding steps. At each step, the model produces a logit vector containing an unnormalized score for every candidate next token in its vocabulary, conditioned on the video, the query, and the tokens generated so far. Applying softmax to each logit vector yields the corresponding next-token probability distribution, which describes the model's uncertainty when generating that part of the timestamp.

We estimate confidence from these distributions at the steps that generate the start and end timestamps, excluding surrounding text such as "from" and "to" and the unit marker "s". For each boundary $b \in \{\mathrm{start}, \mathrm{end}\}$, let T_b denote the number of tokens in its timestamp and let $k = 1, \ldots, T_b$ index their decoding steps in generation order. Let $\mathbf{z}_{b,k}$ denote the logit vector at step k and $\mathbf{p}_{b,k} = \mathrm{softmax}(\mathbf{z}_{b,k})$ its probability distribution. We quantify the uncertainty at this step by Shannon entropy:

$$H(\mathbf{p}_{b,k}) = -\sum_{i=1}^{V} p_{b,k,i} \log p_{b,k,i}, \tag{3}$$

where V is the model's vocabulary size, $p_{b,k,i}$ is the probability of candidate token i, and $\log$ denotes the natural logarithm. The entropy is computed over all candidate tokens at that step; lower entropy indicates that the model's next-token distribution is more concentrated.

To emphasize earlier decisions in autoregressive timestamp generation, we aggregate the token entropies for each boundary using exponentially decreasing weights:

$$s_b = \frac{\sum_{k=1}^{T_b} 2^{-(k-1)} H(\mathbf{p}_{b,k})}{\sum_{k=1}^{T_b} 2^{-(k-1)}}. \tag{4}$$

We then average the two boundary scores to obtain the sample-level uncertainty score:

$$S_{\mathrm{conf}} = \frac{s_{\mathrm{start}} + s_{\mathrm{end}}}{2}. \tag{5}$$

Although denoted by S_{conf}, this score measures uncertainty: lower values indicate lower weighted token entropy and thus stronger expressed confidence in the generated timestamps. It serves as a confidence proxy, not a measure of localization accuracy.

Because entropy scales vary across datasets and model families, we choose the threshold δ_{conf} from the empirical distribution of confidence scores. We then divide the training samples into high- and low-confidence groups:

$$\mathcal{D}^{+}_{\mathrm{conf}} = \{x_i : S_{\mathrm{conf}}(x_i) \le \delta_{\mathrm{conf}}\}, \qquad \mathcal{D}^{-}_{\mathrm{conf}} = \{x_i : S_{\mathrm{conf}}(x_i) > \delta_{\mathrm{conf}}\}. \tag{6}$$

Here, x_i denotes the i-th video–query sample. Since lower entropy indicates higher confidence, $\mathcal{D}^{+}_{\mathrm{conf}}$ contains the more confident samples, while $\mathcal{D}^{-}_{\mathrm{conf}}$ contains the less confident ones.

Confidence intervention. We select a prefill layer ℓ_j using a few-shot layer-wise ablation. Let $\mathbf{h}_{\mathrm{prefill}}(x_i)$ denote the prefill activation extracted at layer ℓ_j for sample x_i. We compute the confidence direction by contrasting the mean activations of the two subsets:

$$\mathbf{v}_{\mathrm{conf}} = \mathbb{E}_{x_i \in \mathcal{D}^{+}_{\mathrm{conf}}} \left[\mathbf{h}_{\mathrm{prefill}}(x_i)\right] - \mathbb{E}_{x_i \in \mathcal{D}^{-}_{\mathrm{conf}}} \left[\mathbf{h}_{\mathrm{prefill}}(x_i)\right]. \tag{7}$$

We then inject this direction into the prefill activation at the same layer:

$$\mathbf{h}'_{\text{prefill}} = \mathbf{h}_{\text{prefill}} + \alpha_{\text{conf}}\mathbf{v}_{\text{conf}}, \tag{8}$$

where α_{conf} controls the intervention strength.

This intervention serves a different purpose from a conventional test-time steering vector. During prefill, the model processes the video and query to form the initial hidden representation before generating any output token. Modifying this representation influences how the model subsequently generates the start and end timestamps. We therefore call the resulting hidden states *confidence-conditioned* (or *internalized*) representations. Confidence is used to shape these representations before the localization direction is constructed by subtracting the mean activation of low-IoU samples from that of high-IoU samples, rather than being introduced as a separate steering signal during inference. This design is central to C2I-CasITI-VTG: confidence steering is used to condition the representation space from which the localization direction is constructed.

After the confidence intervention, we execute the model again and collect the d-dimensional activation vector of the last token at a selected decoding layer ℓ_k. To suppress fluctuations tied to individual interval-token steps, we then average these vectors element-wise over the T decoding steps across the whole decoding process:

$$\overline{\mathbf{h}}_{\text{decode}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_{\text{decode}}(t), \tag{9}$$

where T is the length of the whole decoding process (e.g., generating "from $3.6\,\text{s}$ to $9.8\,\text{s}$"). The averaged activation, rather than the original decoding states, is used in Stage II.

3.3 Stage II: IoU Direction Estimation from High-Confidence Samples

High-confidence Sample Filtering. The second stage focuses on localization quality using the representations obtained after the confidence intervention. We first recompute the confidence scores for the predictions generated with this intervention and retain only samples that meet the adaptive high-confidence threshold:

$$\mathcal{D}_{\text{gate}} = \{x_i : S^{(2)}_{\text{conf}}(x_i) \le \delta^{(2)}_{\text{conf}}\}, \tag{10}$$

where x_i denotes the i-th video–query sample and (2) denotes quantities measured after Stage I.

IoU Direction Estimation. Within the gated set, we partition samples by their ground-truth overlap. For an IoU threshold δ_{IoU},

$$\mathcal{D}^{+}_{\text{IoU}} = \{x_i \in \mathcal{D}_{\text{gate}} : \text{IoU}(\hat{\tau}_i, \tau_i) \ge \delta_{\text{IoU}}\}, \quad \mathcal{D}^{-}_{\text{IoU}} = \{x_i \in \mathcal{D}_{\text{gate}} : \text{IoU}(\hat{\tau}_i, \tau_i) < \delta_{\text{IoU}}\}. \tag{11}$$

Using the confidence-conditioned activations from Eq. 9, we construct the final localization direction

$$\mathbf{v}_{\text{IoU}} = \mathbb{E}_{x_i \in \mathcal{D}^{+}_{\text{IoU}}}\left[\overline{\mathbf{h}}_{\text{decode}}(x_i)\right] - \mathbb{E}_{x_i \in \mathcal{D}^{-}_{\text{IoU}}}\left[\overline{\mathbf{h}}_{\text{decode}}(x_i)\right]. \tag{12}$$

where $+$ and $-$ correspond to the two subsets in Eq. 11.

Although Eq. 12 uses the same difference of mean activations as standard ITI, it compares high- and low-IoU samples within a high-confidence subset whose activations have already been modified by the confidence intervention. Confidence thus serves two purposes: it shapes the hidden representations and determines which samples are retained for constructing the localization direction. The two stages operate sequentially: Stage I modifies the representations from which Stage II constructs the localization direction, rather than providing a separate vector to be applied alongside it.

3.4 Inference-Time Deployment

At test time, only the constructed IoU direction is injected during decoding:

$$\mathbf{h}'_{\text{decode}} = \mathbf{h}_{\text{decode}} + \alpha_{\text{IoU}}\mathbf{v}_{\text{IoU}}. \tag{13}$$

The backbone remains frozen, and only the IoU direction is deployed at test time.

Table 1: Comparison with existing methods on Charades-STA, ActivityNet-Captions, and QVHighlights. All values are percentages. R1 denotes Recall@1. Bold denotes the highest reported value in each metric column. A dash indicates an unreported result. Existing-method results are taken from their respective papers.

Method	Charades-STA				ActivityNet-Captions				QVHighlights			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Fine-tuned MLLM-based methods												
VTG-LLM (Guo et al., 2025)	52.0	33.8	15.7	-	-	-	-	-	-	-	-	-
TimeChat (Ren et al., 2024)	47.7	22.9	12.5	30.6	30.2	16.9	8.2	21.8	-	-	-	-
VTimeLLM (Huang et al., 2024)	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4	-	-	-	-
Momentor (Qian et al., 2024)	42.9	23.0	12.4	29.3	42.6	26.6	11.6	28.5	-	-	-	-
HawkEye (Wang et al., 2024)	50.6	31.4	14.5	33.7	49.1	29.3	10.7	32.7	-	-	-	-
TimeSuite (Zeng et al., 2025)	69.9	48.7	24.0	-	-	-	-	-	-	-	-	-
VideoChat-Flash (Li et al., 2026)	74.5	53.1	27.6	-	-	-	-	-	-	-	-	-
Time-R1 (Wang et al., 2025a)	78.1	60.8	35.3	-	**58.6**	**39.0**	21.4	-	-	-	-	-
Training-free MLLM-based methods												
Moment-GPT (Xu et al., 2025)	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8	-	58.9	38.6	-
NumPro (Wu et al., 2025)	60.7	36.8	15.9	38.5	44.2	26.4	14.4	31.3	-	-	-	-
TFVTG (Zheng et al., 2025)	67.0	50.0	24.3	44.5	49.3	27.0	13.4	34.1	-	-	-	-
T2SGrid (Guo et al., 2026)	71.4	47.0	20.7	44.9	44.8	26.8	16.1	32.5	-	-	-	-
UG-Ground (Kim et al., 2026b)	74.7	52.5	32.6	51.0	54.9	30.8	16.4	38.5	-	-	-	-
Frozen Backbone Baselines and Our Method												
InternVideo2.5-Chat-8B (Wang et al., 2025b)	54.89	30.48	11.24	34.25	22.46	11.46	5.05	17.15	47.55	28.97	13.61	33.09
w/ **C2I-CasITI-VTG**$_{\triangle}$	58.26$_{+3.37}$	33.27$_{+2.79}$	11.97$_{+0.73}$	36.06$_{+1.81}$	26.42$_{+3.96}$	14.14$_{+2.68}$	6.54$_{+1.49}$	19.51$_{+2.36}$	54.97$_{+7.42}$	36.90$_{+7.93}$	18.52$_{+4.91}$	38.14$_{+5.05}$
InternVL3-8B (Zhu et al., 2025)	33.71	12.90	5.40	25.34	21.59	11.31	6.45	17.69	51.29	32.19	19.61	38.06
w/ **C2I-CasITI-VTG**$_{\triangle}$	64.49$_{+30.78}$	34.57$_{+21.67}$	12.18$_{+6.78}$	38.91$_{+13.57}$	42.32$_{+20.73}$	24.41$_{+13.10}$	12.30$_{+5.85}$	29.59$_{+11.90}$	66.84$_{+15.55}$	42.58$_{+10.39}$	25.10$_{+5.49}$	45.91$_{+7.85}$
Qwen3-VL-8B (Bai et al., 2025)	69.60	45.13	19.01	43.92	49.08	32.95	20.03	35.77	67.74	54.97	42.32	53.09
w/ **C2I-CasITI-VTG**$_{\triangle}$	75.03$_{+5.43}$	52.26$_{+7.13}$	23.39$_{+4.38}$	48.03$_{+4.11}$	52.16$_{+3.08}$	35.23$_{+2.28}$	21.03$_{+1.00}$	37.69$_{+1.92}$	68.45$_{+0.71}$	56.13$_{+1.16}$	42.90$_{+0.58}$	53.72$_{+0.63}$
Qwen3.5-9B (Qwen Team, 2026)	79.70	62.63	32.72	53.93	53.02	36.39	22.56	38.83	79.55	67.03	55.35	64.14
w/ **C2I-CasITI-VTG**$_{\triangle}$	**82.12**$_{+2.42}$	**67.15**$_{+4.52}$	**41.16**$_{+8.44}$	**57.47**$_{+3.54}$	55.56$_{+2.54}$	37.86$_{+1.47}$	**23.65**$_{+1.09}$	**40.19**$_{+1.36}$	**81.87**$_{+2.32}$	**69.29**$_{+2.26}$	**57.16**$_{+1.81}$	**66.10**$_{+1.96}$

4 EXPERIMENTS

4.1 DATASETS AND METRICS

We use three standard video temporal grounding benchmarks. Charades-STA contains short, predominantly indoor videos paired with fine-grained descriptions of daily activities. ActivityNet-Captions spans more varied topics and durations, increasing the temporal search space. QVHighlights covers diverse video domains and evaluates query-conditioned localization of salient moments. We follow each benchmark's official split and evaluation protocol.

4.2 IMPLEMENTATION DETAILS

Our primary backbone is Qwen3.5-9B. We additionally evaluate Qwen3-VL-8B, InternVideo2.5-Chat-8B, and InternVL3-8B to assess generalization across model families. All backbone parameters remain frozen. Videos are sampled at 2 frames per second, and all models use the same prompt and greedy decoding strategy. Intervention vectors are constructed from Charades-STA training examples. For Qwen3.5-9B, guided by the median of the empirical entropy distribution, we set the lower and upper entropy thresholds to 0.7 and 1.7, respectively, retaining up to 300 representative samples per group. In Stage II, we set $\delta_{\text{conf}}^{(2)} = 1.0$ and retain samples with $S_{\text{conf}}^{(2)} \leq 1.0$. These samples are further divided using $\delta_{\text{IoU}} = 0.5$ to construct the localization direction. Both interventions are applied at layer 18, with $\alpha_{\text{conf}} = \alpha_{\text{IoU}} = 15$. The localization vectors constructed on Charades-STA are transferred unchanged to ActivityNet-Captions and QVHighlights. Baseline and intervened models share identical preprocessing, prompting, decoding, and evaluation settings. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.3 MAIN RESULTS

Table 1 compares C2I-CasITI-VTG with representative fine-tuned and training-free MLLM-based methods. Existing methods are listed above the frozen-backbone experiments, while our results appear in the lower block. ActivityNet-Captions and QVHighlights additionally evaluate cross-dataset transfer because vectors are constructed only on Charades-STA.

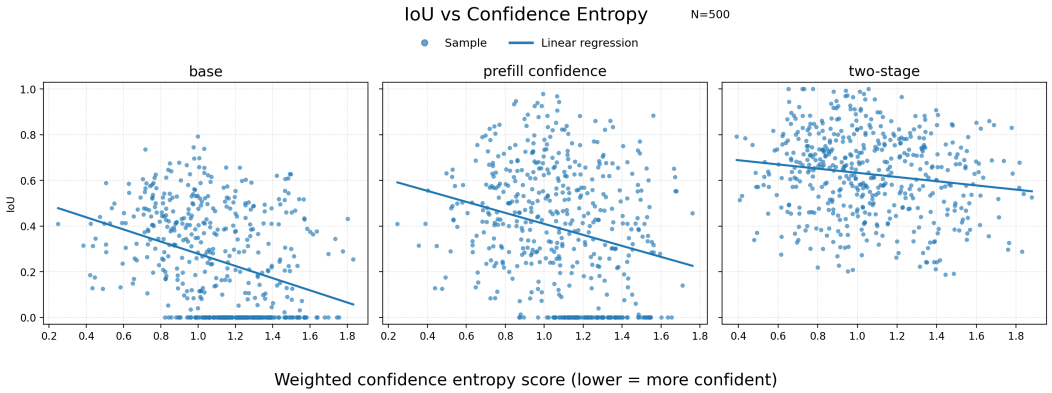


Figure 3: Relationship between confidence entropy and IoU across the base, confidence-prefill, and full two-stage settings. Lower entropy indicates higher confidence.

On Charades-STA, C2I-CasITI-VTG raises Qwen3.5-9B mIoU from 53.93 to 57.47 (an absolute gain of 3.54 points). R1@0.7 increases from 32.72 to 41.16, while the lower thresholds improve as well, indicating gains in both average overlap and strict interval precision. The same IoU vector transfers without additional tuning: mIoU increases from 38.83 to 40.19 on ActivityNet-Captions and from 64.14 to 66.10 on QVHighlights. The positive transfer gains indicate that the constructed direction captures temporal-localization behavior beyond a single dataset's statistics. Across the other frozen MLLM backbones evaluated on Charades-STA, mIoU increases from 43.92 to 48.03 for Qwen3-VL-8B, from 34.25 to 36.06 for InternVideo2.5-Chat-8B, and from 25.34 to 38.91 for InternVL3-8B. Similar gains are also observed on both transfer benchmarks, indicating that the method generalizes across different MLLM architectures.

4.4 Ablation Studies

We conduct ablation studies on Charades-STA to analyze the contribution of each component. We evaluate the role of the two intervention objectives, transferability across backbones, and the effect of restricting second-stage vector estimation to high-confidence samples.

Confidence–localization relationship. Our cascade is motivated by the hypothesis that temporal localization benefits when high expressed confidence is associated with accurate predictions. Fig. 3 traces this relationship across the three stages of our construction. In the base setting, confidence and localization quality are poorly aligned: predictions with similar entropy can exhibit substantially different IoU, including overconfident predictions with poor temporal overlap. The prefill confidence intervention reduces confidence-related uncertainty but does not by itself substantially improve localization, consistent with its role as a conditioning operation rather than a localization objective. After the IoU direction is constructed from the confidence-conditioned hidden-state, the two-stage setting shifts predictions toward the desirable state of lower entropy and higher IoU.

Effect of the intervention objectives. We compare pure IoU intervention, pure confidence intervention, single-stage high-confidence IoU intervention, and the full cascaded design. The pure IoU intervention improves Qwen3.5-9B from 53.93 to 56.13 mIoU, demonstrating that the localization direction is useful. Applying only the confidence intervention gives 44.66 mIoU, confirming that confidence is an auxiliary signal for conditioning the representation rather than a direct localization objective. Restricting IoU-vector construction to high-confidence samples raises mIoU to 56.44, while the complete cascade achieves the best result: 57.47 mIoU and 41.16 R1@0.7. Relative to the baseline, these are gains of 3.54 mIoU points and 8.44 R1@0.7 points.

Transferability across backbones. We apply the same intervention construction and evaluation protocol to Qwen3-VL-8B and InternVL3-8B. As shown in Table 3, the full method consistently improves both backbones. The largest gain is observed for InternVL3-8B, whose mIoU increases from 25.34 to 38.91. Qwen3-VL-8B gains 4.11 mIoU points. The single-stage high-confidence IoU variant is weaker than the full cascade, supporting the role of confidence internalization in constructing the localization direction.

Table 2: Ablation of intervention objectives on Charades-STA using Qwen3.5-9B. All values are percentages.

Method	R1@0.3	R1@0.5	R1@0.7	mIoU
Qwen3.5-9B baseline	79.70	62.63	32.72	53.93
Pure IoU intervention	80.91	65.67	38.12	56.13
Pure confidence intervention	71.91	44.92	18.12	44.66
Single-stage high-confidence IoU	81.36	65.90	38.76	56.44
C2I-CasITI-VTG	**82.12**	**67.15**	**41.16**	**57.47**

Table 3: Backbone transferability on Charades-STA. The intervention vector is constructed without updating backbone parameters.

Backbone	Method	R1@0.3	R1@0.5	R1@0.7	mIoU
InternVL3-8B	Baseline	33.71	12.90	5.40	25.34
	Only high-confidence IoU	58.44	29.54	10.81	36.33
	C2I-CasITI-VTG	**64.49**	**34.57**	**12.18**	**38.91**
Qwen3-VL-8B	Baseline	69.60	45.13	19.01	43.92
	Only high-confidence IoU	70.67	46.37	20.51	44.88
	C2I-CasITI-VTG	**75.03**	**52.26**	**23.39**	**48.03**
Qwen3.5-9B	Baseline	79.70	62.63	32.72	53.93
	Only high-confidence IoU	81.36	65.90	38.76	56.44
	C2I-CasITI-VTG	**82.12**	**67.15**	**41.16**	**57.47**

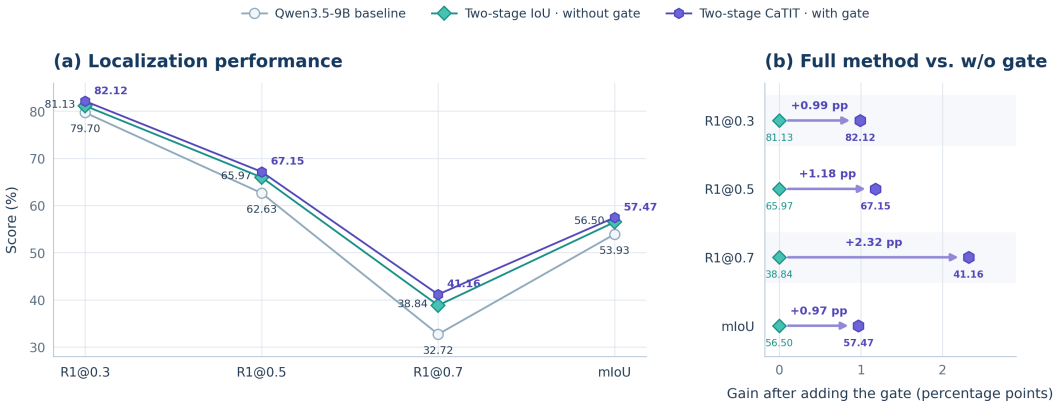


Figure 4: Ablation of the high-confidence gate on Charades-STA with Qwen3.5-9B. (a) Localization scores of the baseline, the two-stage variant without high-confidence gate, and the full method with the gate. (b) Absolute gains of the full method over the variant without the high-confidence gate.

Importance of high-confidence filtering in Stage II. As shown in Fig. 4(a), both two-stage variants outperform the baseline, while the high-confidence gate further improves all four metrics. Compared with the variant without the gate, the full method consistently improves all four evaluation metrics (Fig. 4(b)). The largest gain occurs at the strictest IoU threshold: R1@0.7 increases from 38.84 to 41.16, while mIoU rises from 56.50 to 57.47. These results support the use of high-confidence filtering when constructing the second-stage localization direction from confidence-conditioned hidden-state.

Overall, the ablations support the proposed cascaded construction. Confidence intervention is useful when it internalizes a behavioral prior into the activation space, after which the localization vector is constructed from high-confidence internalized activations. This two-step procedure is stronger than direct IoU intervention or single-stage high-confidence vector construction while retaining single-vector deployment at inference time.

5 Conclusion

We study the confidence–IoU mismatch in video temporal grounding, where predictions with high localization quality may still exhibit uncertain interval behavior, making direct IoU-based activation steering noisy. We address this problem with C2I-CasITI-VTG, which first applies confidence intervention during prefill to alter a confidence-conditioned hidden-state space and then constructs a localization direction from high-confidence decoding activations. The resulting method deploys only a single IoU direction at inference time, without altering the backbone parameters or target-dataset. C2I-CasITI-VTG consistently improves localization on multiple MLLMs across Charades-STA, ActivityNet-Captions, and QVHighlights. These results indicate that cascaded activation steering can coordinate dependent behavioral objectives effectively. Future work could explore more reliable confidence signals, automatic layer and threshold selection, and more efficient cascaded intervention construction across broader domains.

6 STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, draft and reorganize portions of the manuscript, assist with language translation and clarification, format LaTeX equations and mathematical notation, and support the retrieval and discovery of potentially relevant literature. The tools also assisted in wording descriptions of the method and experimental observations based on content and results provided by the authors.

We did not use generative AI tools to originate the core research idea, propose or refine research hypotheses, design the methodology or experiments, implement the method, generate or modify datasets, produce experimental results, execute experiments, or formulate or prove mathematical claims. All references identified with AI assistance were manually checked against the original publications. All AI-assisted text, equations, notation, and structural suggestions were reviewed, revised, and verified by the authors, and all reported results and claims were checked against the actual experimental outputs. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human-subject studies or collection of private or sensitive data. All experiments are conducted on publicly available academic benchmarks and pretrained models for research purposes. We do not release any new dataset or personal information. As with other learning-based systems, the reported results may reflect biases present in the underlying datasets and pretrained models. We identify no additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

The vector-construction procedure is described in Section 3. Experimental settings for our primary backbone, Qwen3.5-9B, are provided in Section 4.2, and the prompt template is included in Appendix A.1. All experiments use publicly available benchmarks and their official evaluation protocols. The code will be released at a later stage.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, and Ziyan Wu. Seq2Time: Sequential Knowledge Transfer for Video LLM Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13766–13775, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Seq2Time_Sequential_Knowledge_Transfer_for_Video_LLM_Temporal_Grounding_CVPR_2025_paper.html.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Jia Jim Deryl Chua, Lee Onn Mak, Gee Wah Ng, and Kezhi Mao. Fine-Grained Activation Steering: Steering Less, Achieving More. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/423d0909791493b7c10916fd328c2913-Abstract-Conference.html.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563.

Chaohong Guo, Yihan He, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. T2SGrid: Temporal-to-Spatial Gridification for Video Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3443–3454, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Guo_T2SGrid_Temporal-to-Spatial_Gridification_for_Video_Temporal_Grounding_CVPR_2026_paper.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. URL https://arxiv.org/abs/1706.04599.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. VTG-LLM: Integrating timestamp knowledge into video LLMs for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3302–3310, 2025. doi: 10.1609/aaai.v39i3.32341. URL https://doi.org/10.1609/aaai.v39i3.32341.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1380–1390, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1168. URL https://aclanthology.org/D18-1168/.

Jian Hu, Zixu Cheng, Shaogang Gong, Isabel Guan, Jianye Hao, Jun Wang, and Kun Shao. Uncertainty-quantified Rollout Policy Adaptation for Unlabelled Cross-domain Temporal Grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://nips.cc/virtual/2025/poster/118007.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

Sanghwan Kim, Rui Xiao, Stephan Alaniz, Yongqin Xian, and Zeynep Akata. Training-free Uncertainty Guidance for Complex Visual Tasks with MLLMs. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2510.00705. Accepted at ECCV 2026; proceedings forthcoming.

Sanghwan Kim, Rui Xiao, Stephan Alaniz, Yongqin Xian, and Zeynep Akata. Training-free uncertainty guidance for complex visual tasks with mllms, 2026b. URL https://arxiv.org/abs/2510.00705.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b14d7175755b180dc2163e15e3110cb6-Paper-Conference.pdf.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. TempSamp-R1: Effective Temporal Sampling with Reinforcement Fine-Tuning for Video LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/39da069a3e1baa8139b62d6603d37979-Abstract-Conference.html.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal Video Temporal Grounding with Generative Multi-modal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5d2e24df9cfaad3189833b819c40b392-Abstract-Conference.html.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2024. URL https://arxiv.org/abs/2306.05424.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation Steering with a Feedback Controller. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa5617c176e76fee83f3f9947fdf9f3f-Abstract-Conference.html.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1007. URL https://aclanthology.org/2025.acl-long.1007/.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41340–41356. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/qian24a.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_TimeChat_A_Time-sensitive_Multimodal_Large_Language_Model_for_Long_Video_CVPR_2024_paper.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7801b29c93b599b8d0c44138596bdeed-Abstract-Conference.html.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025b. URL https://arxiv.org/abs/2501.12386.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024. URL https://arxiv.org/abs/2403.10228.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A Survey on Video Temporal Grounding With Multimodal Large Language Model . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(02):1521–1541, February 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3615586. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3615586.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal Grounding Videos like Flipping Manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Number_it_Temporal_Grounding_Videos_like_Flipping_Manga_CVPR_2025_paper.html.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. HiTeA: Hierarchical Temporal Alignment for Training-Free Long-Video Temporal Grounding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/50e3c627697a456e29ec797c36621516-Abstract-Conference.html.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models, 2025. URL https://arxiv.org/abs/2501.07972.

Zaiquan Yang, Yuhao Liu, Gerhard Hancke, and Rynson W. H. Lau. Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/poster/119064.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic Multimodal Activation Steering for Hallucination Mitigation in Large Vision-Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0480adaf62a918405a5e3b1031e0c056-Abstract-Conference.html.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5e8309c9ca683e11672e3dbcd4b87776-Paper-Conference.pdf.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020a. URL https://arxiv.org/abs/2004.13931.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. VTimeCoT: Thinking by Drawing for Video Temporal Grounding and Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24203–24213, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_VTimeCoT_Thinking_by_Drawing_for_Video_Temporal_Grounding_and_Reasoning_ICCV_2025_paper.html.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking Video Temporal Grounding with Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b. URL https://arxiv.org/abs/1912.03590.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 20–37, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73007-8. URL https://doi.org/10.1007/978-3-031-73007-8_2.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Appendix

A.1 Prompt and Output Processing

We use a common prompt template for the baseline and all intervention variants:

> To accurately pinpoint the event "[QUERY]" in the video, determine the precise time period of the event. Provide the start and end times in the format: "From x to y."

Here, [QUERY] is replaced with the natural-language query associated with the video.

We parse the first valid temporal range from the generated response. The parser accepts expressions containing "from" or "between" followed by "to" or "and", as well as ranges separated by a dash, comma, or tilde. Timestamps may be expressed directly in seconds or in mm:ss and hh:mm:ss formats. Parsed timestamps are converted to seconds and clipped to the valid video duration. If no valid pair can be extracted, the prediction is set to $[0, 0]$. The same parsing procedure is used for the baseline and all intervention variants.

A.2 Activation Extraction and Layer Selection

Hooks are registered on the attention output of every language-model layer. For layers implemented with linear attention, the corresponding linear-attention output is used.

Our layer-wise experimental results show that interventions at relatively early layers yield poor performance, possibly because they disrupt the ongoing fusion of visual and linguistic information. Performance also declines at later layers, possibly because these layers are closer to the output and

leave insufficient downstream computation for the intervention to propagate. For clarity, we report only the results for layers 16–25, which contain the best-performing region and the selected layer.

For Qwen3.5-9B, we evaluate these candidate layers on the same 150 held-out Charades-STA training samples while keeping the remaining inference settings fixed. We use mIoU as the selection criterion. As shown in Fig. 5, layer 18 achieves the highest mIoU and is used in the subsequent Qwen3.5-9B experiments.

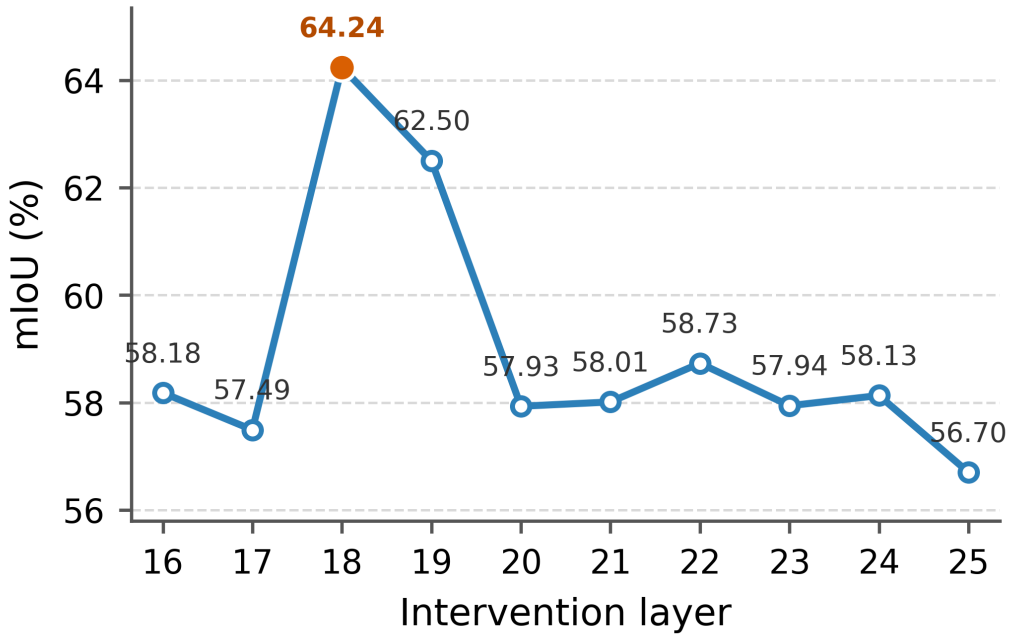


Figure 5: Layer selection for Qwen3.5-9B on 150 held-out Charades-STA samples. Layer 18 achieves the highest mIoU of 64.24.

For Qwen3.5-9B, intervention is applied to the complete 4096-dimensional hidden vector at layer 18, with $\alpha_{\text{conf}} = \alpha_{\text{IoU}} = 15$. Videos are sampled at 2 FPS and decoded greedily. Experiments are conducted on one NVIDIA RTX 3090 GPU. The corresponding layers and intervention strengths for the other backbones are selected on the validation set using the same protocol.